ABSTRACT

A method and apparatus for provisioning bandwidth among a plurality of queues in a switching device are disclosed. The bandwidth provisioning employs a QoS label with a timestamp to determine the queue delay associated with the transmission of the packet from an ingress switching device, e.g. ingress blade, to an egress switching device, e.g. egress blade. The QoS label is applied on the ingress blade and removed and read on the egress blade. Using the timestamp, the switch can identify congested priority queues at the ingress blade and selectively enable fragmentation for purposes of preventing particular packets and or particular flows from consuming disproportionately large system resources including bandwidth necessary to communicate packets through the switch fabric. Some QoS labels may further include instruction sets and a fragmentation identifiers used by the egress blade to reassemble the packet fragments. The QoS label, in cooperation with computational and management resources, provide a feedback mechanism to perform real-time bandwidth provisioning within a switch.